A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Yumiko Maria Bejarano Azogue |
| *Student Number* | 2024144 |
| *Module Title* | Statistical Techniques for Data Analysis |
| *Assessment Title* | CA 1 – Capstone Project Proposal |
| *Assessment Due Date* | 21st April 2024 |
| *Date of Submission* | 21st April 2024 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Contents

[Task 1 - Data Analysis: 3](#_Toc164556920)

[Data Source Overview 3](#_Toc164556921)

[Data Analysis (Characterization of Data) 3](#_Toc164556922)

[Data Description 3](#_Toc164556923)

[Summary Statistics: 4](#_Toc164556924)

[Dependent variable: Salary: 5](#_Toc164556925)

[Pre-processing 5](#_Toc164556926)

[Statistical Analysis 6](#_Toc164556927)

[Conclusion 7](#_Toc164556928)

[Task 2 - Probability (Discrete): 8](#_Toc164556929)

[Question 1 8](#_Toc164556930)

[Question 2 9](#_Toc164556931)

[Task 3 Probability (Continuous): 11](#_Toc164556932)

[Question 11](#_Toc164556933)

[References 13](#_Toc164556934)

# Task 1 - Data Analysis:

## Data Source Overview

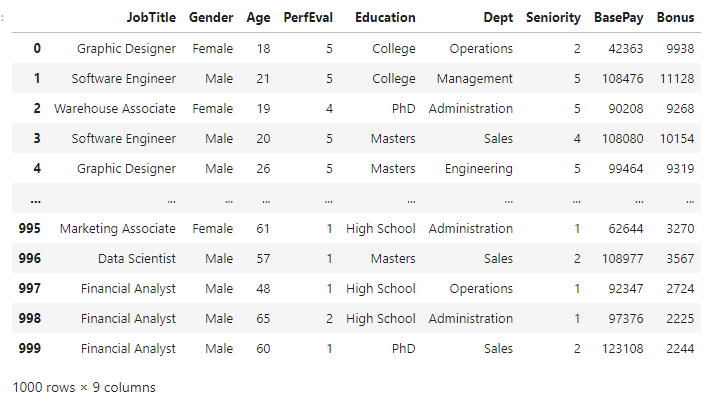
The dataset was obtained from Kaggle and focuses on incomes for various job titles by gender. Below is the link to access the dataset.

<https://www.kaggle.com/datasets/nilimajauhari/glassdoor-analyze-gender-pay-gap>

## Data Analysis (Characterization of Data)

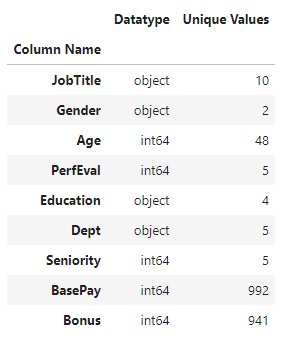
### Data Description

The dataset comprises 1000 entries and includes features such as Job Title, Gender, Age, Performance Evaluation, Education, Department, Seniority, Base Salary, and Bonus.



Dataset size: 9 rows, 8 columns

Variables: It includes nine variables, encompassing ***five numerical and four categorical*** variables.

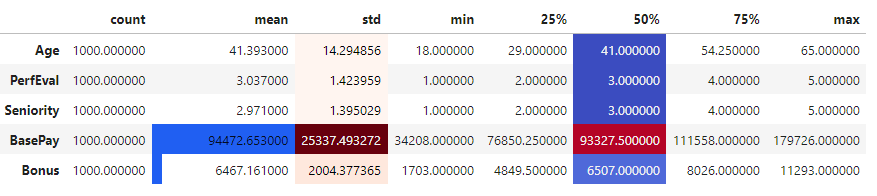
It has nine columns, namely-

* Job Title
* Gender
* Age
* PerfEval - Performance Evaluation Score
* Education
* Dept
* Seniority - Number of years worked.
* BasePay - Annual basic pay in US Dollars.
* Bonus - Annual bonus in US Dollars.

### Summary Statistics:

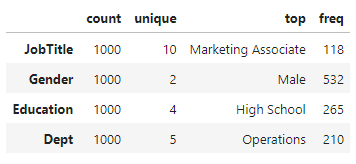
Descriptive statistics for numerical variables in the dataset:

* Age: The average age of individuals is approximately 41 years, with a standard deviation of around 14 years. Age ranges from 18 to 65 years.
* PerfEval: The average performance evaluation score is about 3.04, with scores ranging from 1 to 5.
* Seniority: On average, individuals have a seniority of around 2.97 years, with a range from 1 to 5 years. Half of the individuals have seniority equal to or less than 3 years.
* BasePay: The average base salary is approximately $94,472.65, with salaries ranging from $34,208 to $179,726.
* Bonus: The average bonus is about $6,467.16, with bonuses ranging from $1,703 to $11,293.
* The minimum recorded bonus is $1,703, and the maximum is $11,293.



Frequencies of the categorical variables in our dataset:

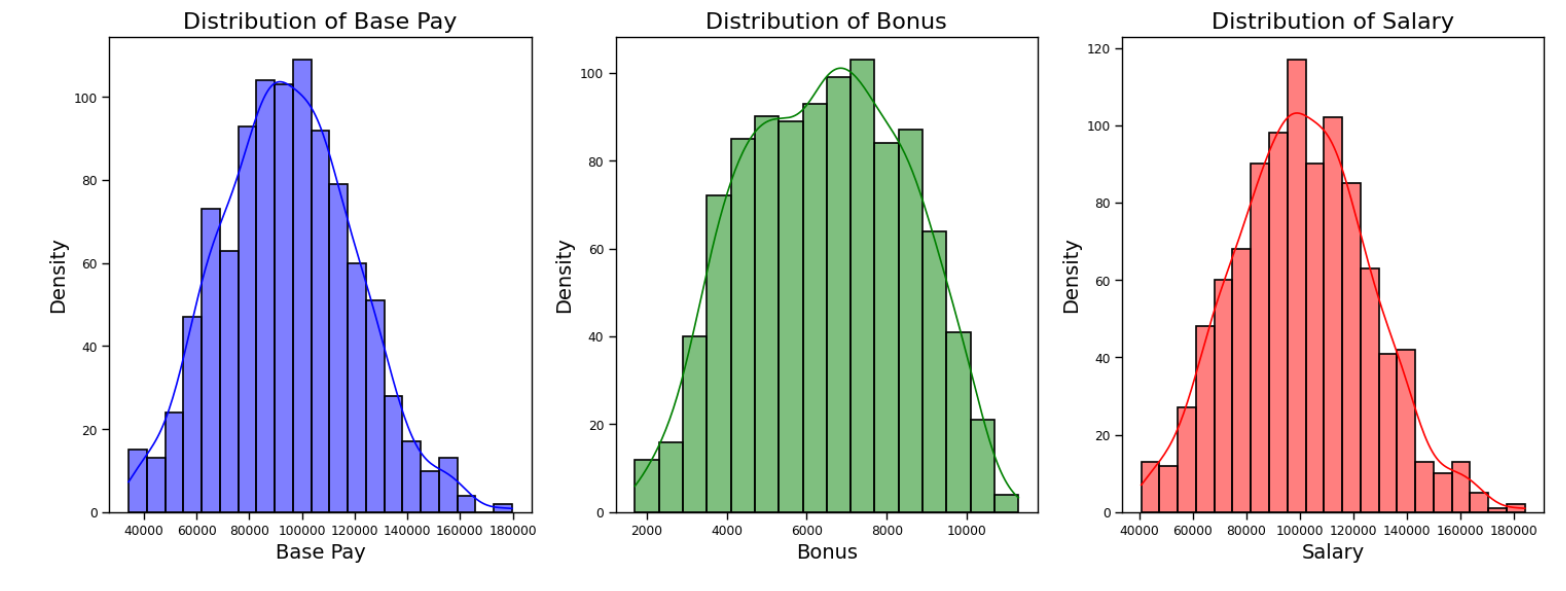
* JobTitle: There are 10 job titles in the dataset, with "Marketing Associate" being the most common, appearing 118 times.
* Gender: There are two categories: "Male" and "Female," with "Male" being the most frequent, totaling 532 records.
* Education: There are four educational levels, with "High School" being the most common, with 265 records.
* Dept: There are five departments, with "Operations" being the most common, with 210 records.

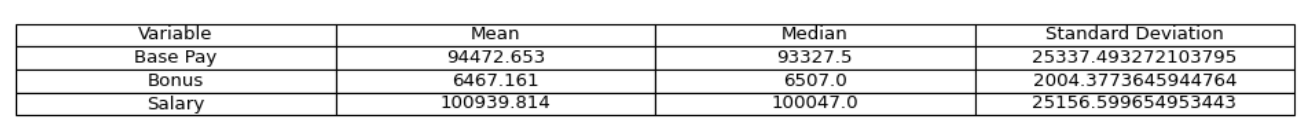


### Salary:

The salary is calculated as sum of the base salary and the yearly bonus.



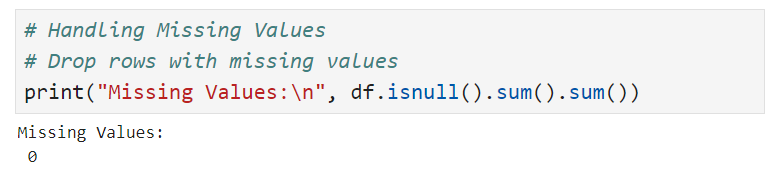




### Pre-processing

#### Handling Missing Values:

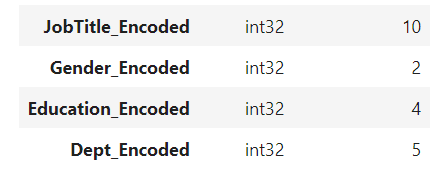
**No missing values** were detected in any of the variables in the dataset, as confirmed during the data exploration phase. Therefore, no correction or removal of missing values is required.



##### Encoding Categorical Variables:

To facilitate the analysis of the dataset, categorical variables were encoded into a numerical format. In our dataset, we identified four categorical variables: "Job Title," "Gender," "Education," and "Department".

The label encoding method was used to transform these variables into numerical values. The results of this encoding are shown at the end, allowing for more effective use of the data in subsequent analyses.



* Normalization/Standardization: Scale numerical features to a similar range to prevent variables with larger magnitudes from dominating the analysis.

## Analyse by visualizing data (Statistical Analysis)

**Mean, Median and Mode**

|  |  |  |
| --- | --- | --- |
| Mean | Median | Mode |
|  |  |  |

In the histograms, we observe the distribution of salaries for all individuals in the dataset, alongside their respective mean, median, and mode values for both men and women.

For the overall dataset:

* Median Salary: $100,047.0
* Mean Salary: $100,939.814
* Mode Salary: $58,373

For men:

* Median Salary: $105,100.5
* Mean Salary: $104,918.68
* Mode Salary: $98,578

For women:

* Median Salary: $96,571.0
* Mean Salary: $96,416.83
* Mode Salary: $83,172

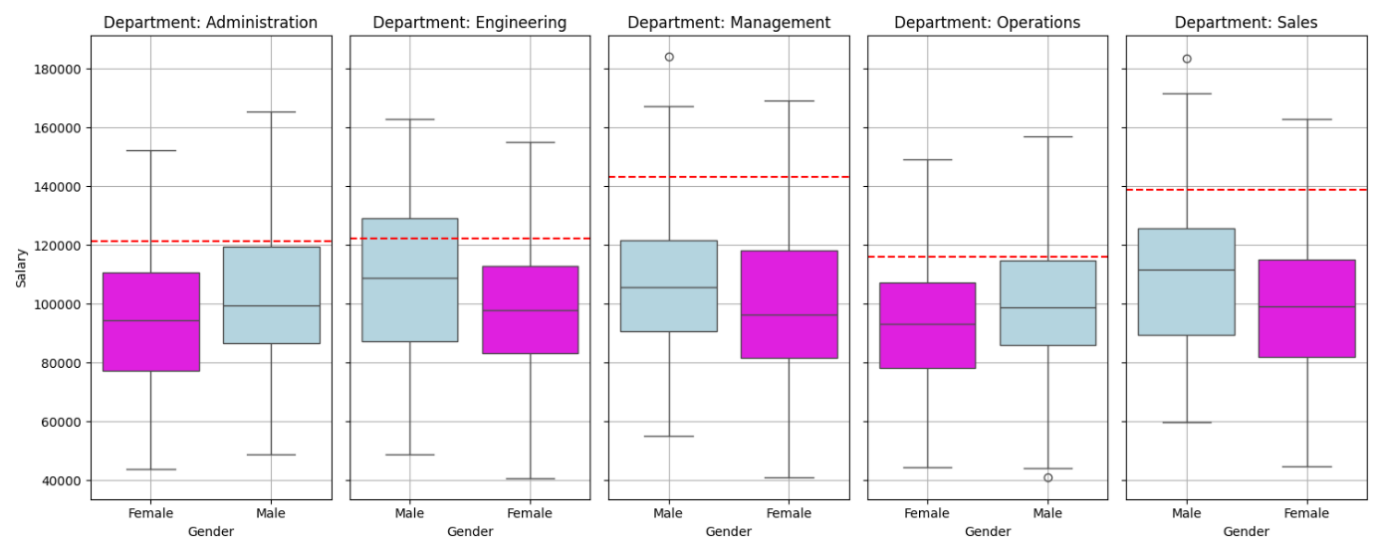
These results provide a comprehensive view of the salary distribution within the dataset. We can observe that the median salary for men is higher than that for women, indicating a potential gender disparity. Similarly, the mode salary for men is higher than that for women, suggesting a concentration of higher salaries among men. The mean salary also reflects this trend, with men having a higher average salary compared to women.

**Range**

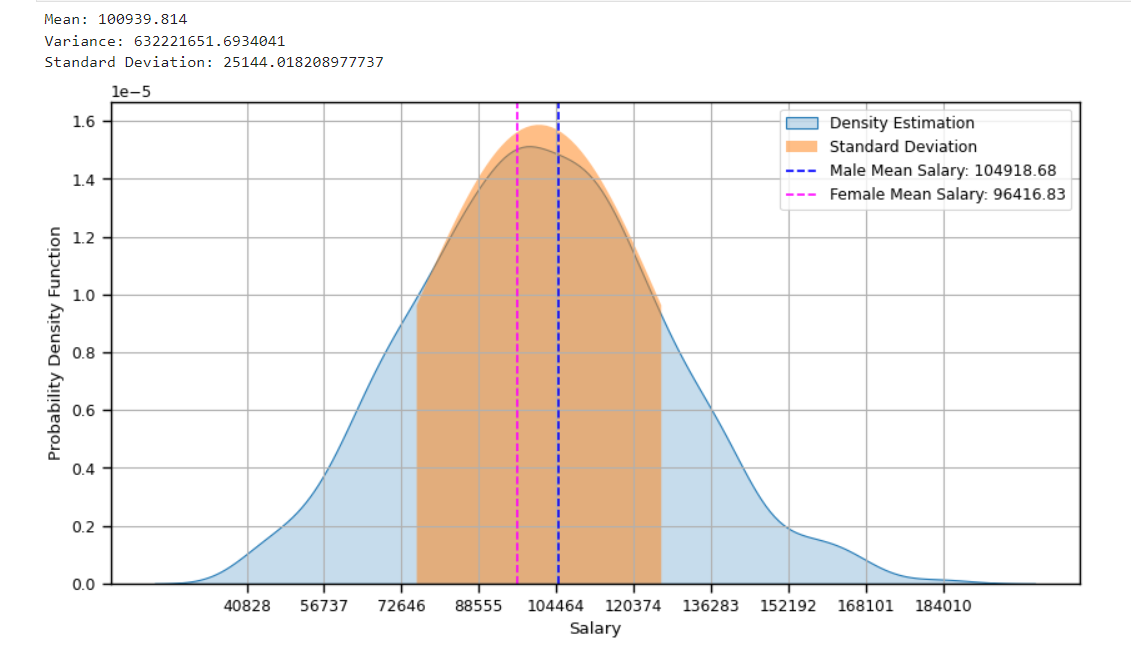
We looked at the salary ranges for men and women in each department. Here's what we found:

* In Administration, salaries vary about the same for both men and women.
* In Engineering, men's salaries vary more than women's.
* In Management, salaries vary similarly for both men and women.
* In Operations, salary differences between men and women are about the same.
* In Sales, the salary ranges are similar for both men and women.

These differences might show us where there could be unequal pay between men and women in some departments.



Standard deviation



## Conclusion

In conclusion, this study sheds …...

# Task 2 - Probability (Discrete):

### Question 1

What is the probability of rolling exactly two 6s in five rolls of a fair die?

**Binomial Distribution:** The Binomial distribution represents the number of successes in a fixed number of independent Bernoulli trials. It describes the number of successes k in n independent experiments, each with a probability p of success. The probability mass function of the Binomial distribution is given by the formula:



​Where:

*n* = number of occurrences of a specific outcome in n trials

*p* = probability of success in a single trial

*k* = number of trials

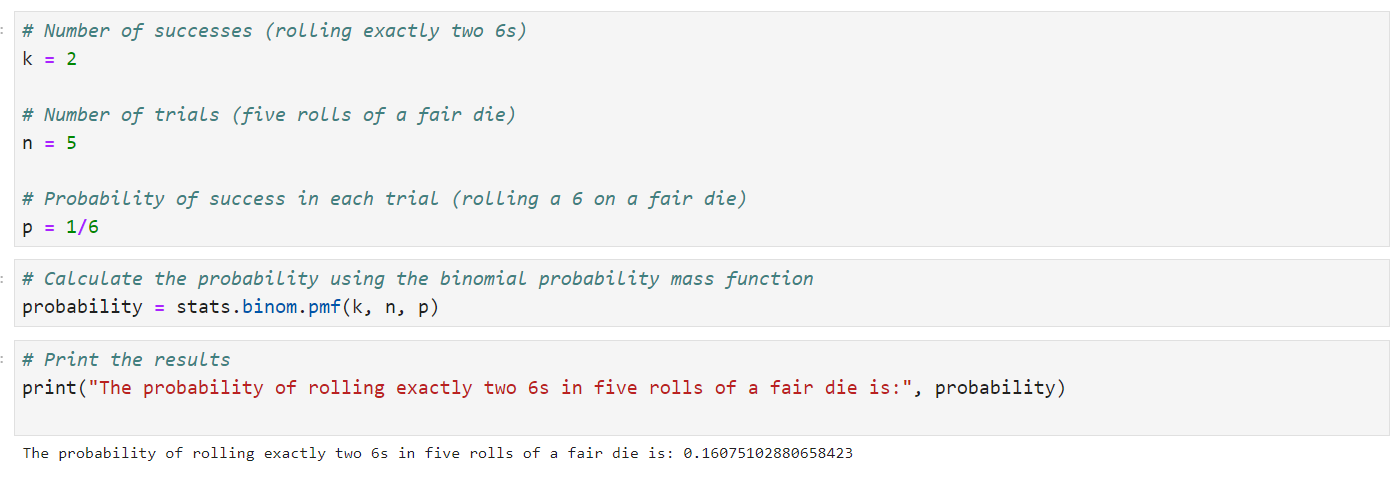
= number of combinations

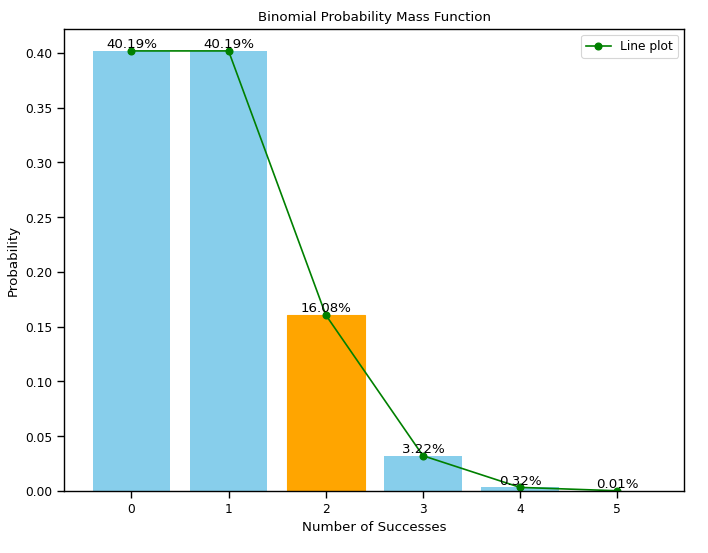
#### **Use the binomial probability**

Substituting the given values:

Binomial Coefficient

So, the probability of rolling exactly two 6s in five rolls of a fair die is approximately 0.160751

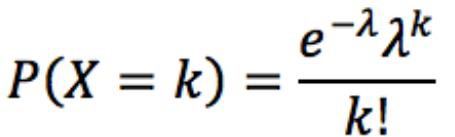




### Question 2

The number of industrial injuries on average per working week in a factory is 0.75. Assuming that the distribution of injuries follows a Poisson distribution, find the probability that in a particular week there will be no more than two accidents

**Poisson Distribution:** The Poisson distribution is used to model the number of events occurring in a fixed interval of time or space, under the assumption that these events occur with a known constant mean rate and are independent of the time since the last event. It is characterized by a single parameter, λ, which represents the average rate of occurrence over a given interval. The probability mass function of the Poisson distribution is given by:



The Poisson cumulative distribution function (CDF) is used because we want to find the probability of up to a certain number of events occurring in a given interval, rather than exactly that number.

Where:

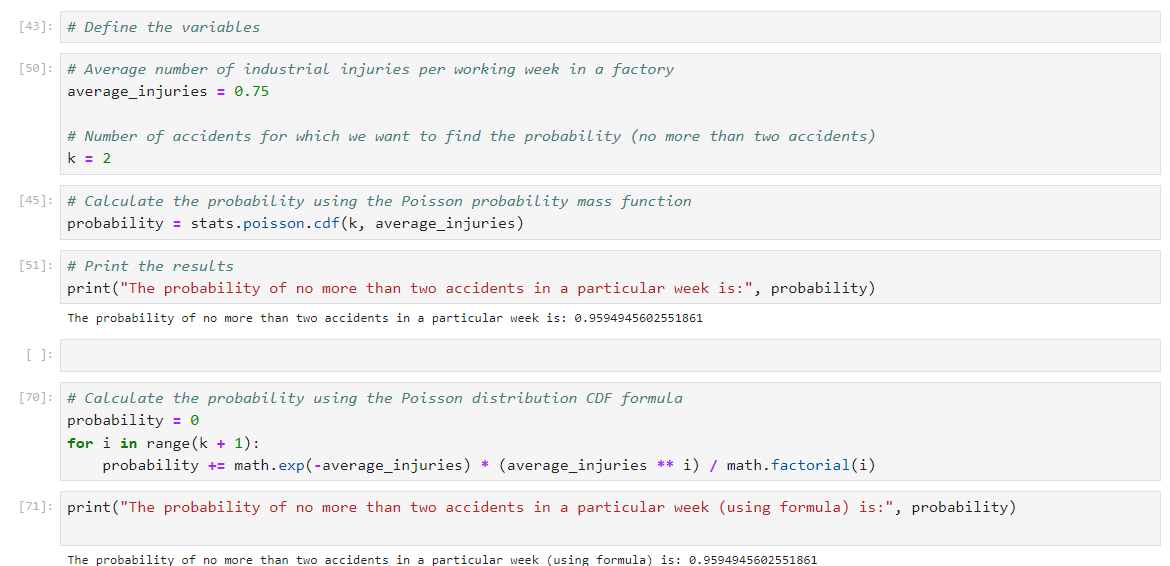
* *k* is the number of occurrences (Poisson random variable)
* λ is the rate of success (Greek letter lambda)
* *e* ≈ 2.71828 (Euler’s number)
* both x and λ are non-negative integers

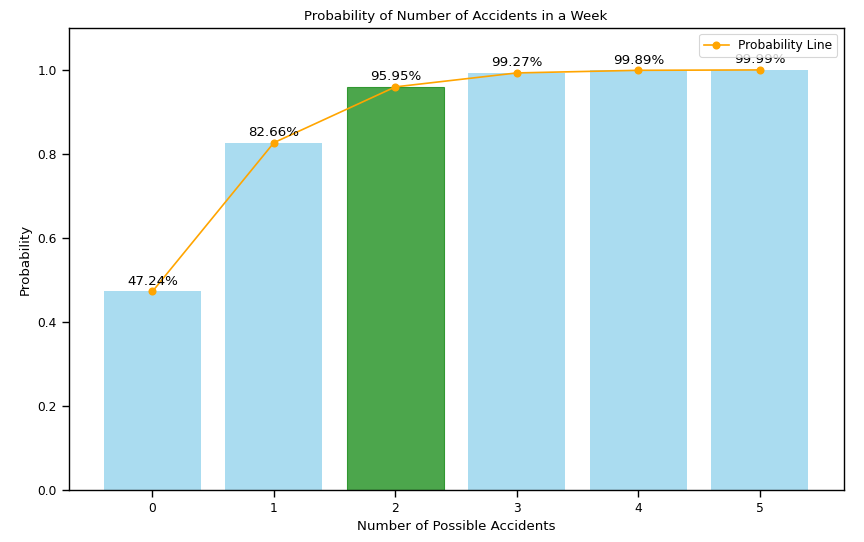
Mean (average) number of accidents per week (λ) = 0.75

#### **Probability using the Poisson cumulative distribution:**

Substituting the given values:

So, the probability that in a particular week there will be no more than two accidents is approximately 0.959494





# Task 3 Probability (Continuous):

### Question

The time a person spends at Dublin Zoo is Normally distributed with a mean of 90 minutes and a standard deviation of 10 minutes.

Using this distribution, answer the following:

* If a visitor is selected at random, find the probability that they will spend at most 85 minutes visiting the zoo.
* If a visitor is selected at random, find the probability that they will spend at least 100 minutes visiting the zoo.
* Given that you know that a particular visitor has spent longer than average visiting the Zoo, what is the probability that they have spent more than 100 minutes there?

Given the mean (μ\\muμ) and standard deviation (σ\\sigmaσ), we can find probabilities using the cumulative distribution function (CDF) of the normal distribution.

* Mean (μ) = 90 minutes
* Standard deviation (σ) = 10 minutes

We'll use the standard normal distribution (with mean μ=0 and standard deviation σ=1) and then adjust for the given mean and standard deviation.

1. *Probability of spending at most 85 minutes:*

Using the standard normal distribution table or a calculator, we find 0.3085 .

1. *Probability of spending at least 100 minutes:*

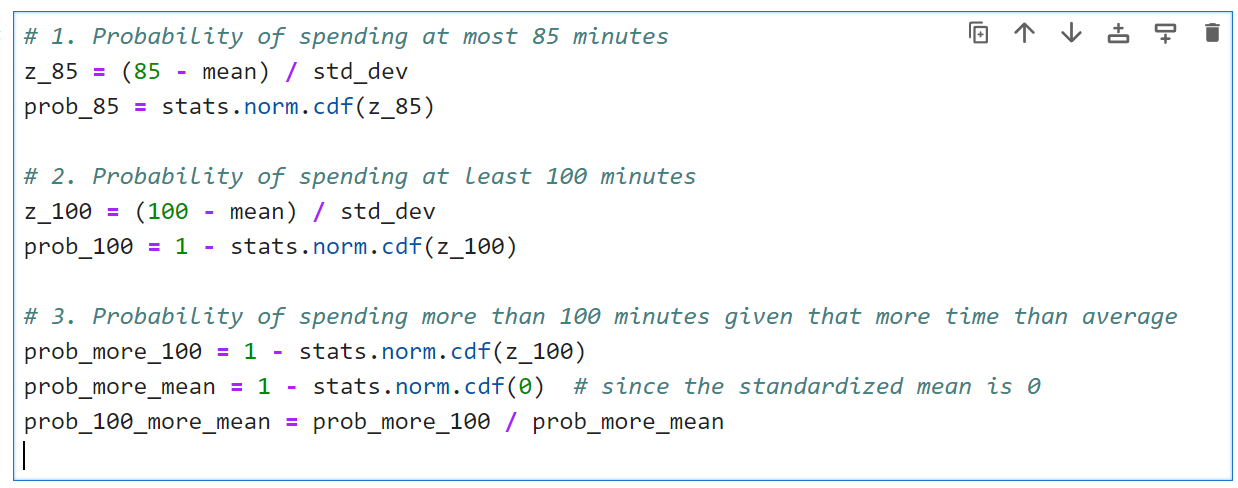
Using the standard normal distribution table or a calculator, we find .

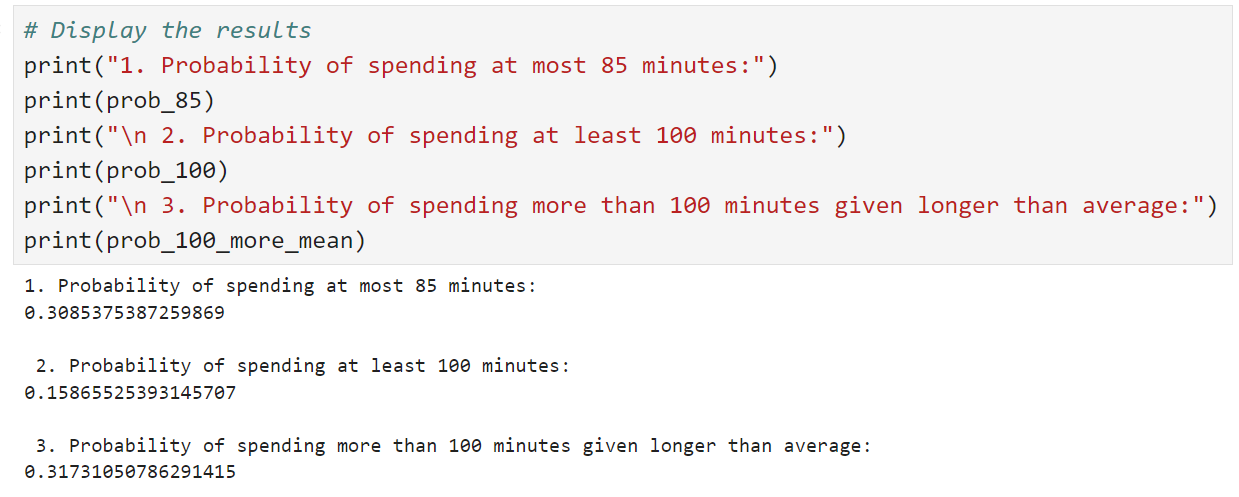
Therefore,

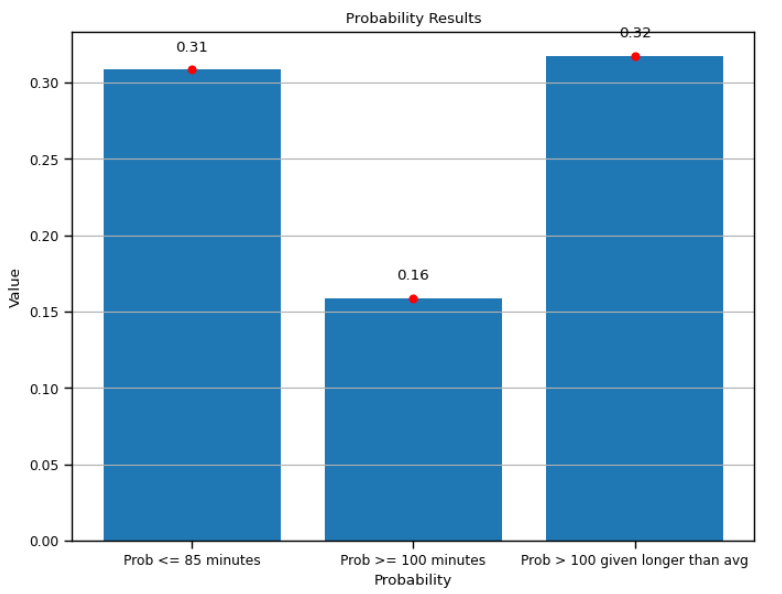
1. *Probability of spending more than 100 minutes given longer than average:*

To find , we use the standard normal distribution table or a calculator to find .

So, the probabilities are:







# References

Bonthu, H. (2021). *Exploring The Different Types Of Probability Distribution Function!* [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/07/probability-types-of-probability-distribution-functions/.

F.M., D., C., K., H.P. , L. and L.E. , M. (2005). *A modern introduction to probability and statistics : understanding why and how*. New York: Springer, Cop.

Keane, J. (n.d.). *Understanding The Data - Irish Gender Pay Gap Portal*. [online] paygap.ie. Available at: https://paygap.ie/understandingTheData [Accessed 16 Apr. 2024].

OpenAI (2024). *ChatGPT*. [online] chat.openai.com. Available at: https://chat.openai.com/ Get instant answers, find inspiration, learn something new.

Wallis, S. (2013). z-squared: *The Origin and Application of. Journal of Quantitative Linguistics*, 20(4), pp.350–378. doi:https://doi.org/10.1080/09296174.2013.830554.

www.kaggle.com. (n.d.). *Glassdoor- Analyze Gender Pay Gap*. [online] Available at: https://www.kaggle.com/datasets/nilimajauhari/glassdoor-analyze-gender-pay-gap.